Text Classification - Amazon Reviews

Poorva Araj
COMP1804 Applied Machine learning
000997514

***Abstract*—In this report, the implementation of a machine learning model, 'Text classification-Amazon Reviews' is described. This machine learning model aims to classify text reviews given by customers to identify the Rating associated with that review on a scale of 1-5. Further based on these reviews product category is identified, whether the product is a musical instrument or a video game. This machine learning model was built using supervised learning algorithms- Logistic regression and Linear support vector classifier. In order to implement the machine learning model, Initially Data quality assessment followed by Exploratory data analysis was carried out. As some data columns are in text format, text Pre-processing is performed in order to transform the data into clean and noiseless format. After text Pre-processing, the data frame is split into train and test data for systematic evaluation. Further, we fit the data into respective machine learning model and make prediction. In order to evaluate the data various evaluation matrices were used to examine reliability of model.**

I. Introduction and related work

In e-commerce businesses it is important to understand the customers and their requirements. However in the large scale businesses a company cannot keep track of thousands of customer's text reviews. In this situation, there should be a model which could transform the long texts reviews into a simple scale rating to understand customer sentiment effortlessly.

In this machine learning model, reviews on the scale of 1-5 and the category of product is predicted based on the customer's comments in textual format. Reviews prediction can be helpful for the customers to identify the quality of product as well as customers can use rating filters to filter out the low rated items and optimize the search experience. Even for businesses identifying the product category and its rating will help to analyse and improve the products which has highest/lowest ratings. Hence, the identification of ratings and its product category can greatly impact businesses and its customers.

This machine learning model uses Natural language processing for text classification. Text classification is a emerging technology which has multiple use cases. It is used in recommendation systems, sentiment analysis, speech recognition, language detection, fraud detection, etc. Thus there exists similar machine learning models based works.

The dataset used for running a machine learning model is consists of 32918 rows and 5 columns. The columns are consists of review Id, text- A customer review in text format, Verified- if the review from the customer is legit, review score- on the scale -1 to 5 and product category- if the product is from video game category or musical instrument. Initially some data quality assessment was performed to gather details of mismatched data types or features, checked the data values whether they follow similar format, etc. Further we did exploratory data analysis where we gathered details of missing values, count of values, deviations in values, and check if the dataset is balanced or imbalanced. In the next step, data Preprocessing was performed. In the preprocessing data points with missing values were dropped. There were 12 rows of 'text' column and 906 rows of 'product category' were having Null value. As this was a textual data filling this null points would have created noise in the data. Hence they were dropped.

In order to make accurate predictions, text preprocessing was performed and Further, Label encoder was applied on product category column to replace categories into numerical format.

To identify the product categories, Binary logistic regression algorithm is used. It uses a logistic function to identify and predict discrete variables and provide binary output variable. As we needed to predict only two discrete categories, whether it is Musical Instrument or video game, Binary classification was a suitable algorithm for the task. In the next task to identify the customer ratings, whether it is 1, 2, 3, 4 or 5 Linear SVC (support vector classifier) algorithm is used. After implementing multiple algorithms like naive-bayes, KNN for predicting ratings, Linear SVC provided more accuracy score. In linearSVC with the one vs All approach rating classes were divided into each 5 classes, 1,2,3,4 and 5. LinearSVC finds a Hyperplane which will divide and distinguish between the classes.

II. Ethical discussion

In this machine learning model, data which is provided must be collected from official amazon websites, from where customers have ordered a product. The data simply contains the review id, review, whether the review is verified or not, review score and product category. This data does not include any kind of personal information of customer like name, credit card information or Date of birth. Apart from that the reviews collected are consented by customers. This ML model aims to provide better features and insights to e-commerce businesses as well as customers. This model helps in technological improvements and effectiveness without affecting any jobs or people's lives.

III. Dataset preparation

In the Data preparation stage, we initially explore the data using methods like .describe(), .shape(), .isnull(), .value-counts(), etc.. Dataset contains 32918 rows and 5 columns. For the task 1, to predict the ratings, we select only two columns for further implementation wise 'Text' and 'review_score'. Because the aim of the project is to predict ratings by the customer's reviews. Further for task 2, 'Product_category' and 'text' are only two columns are selected for further implementation. In the data cleaning stage, initially dataset is checked for any null values. The text column has 12 empty cells, product_category has 906 and review _has 0. The null cells from text category has to be dropped because filling the null text cells could lead to noise creation. Further for the Product_category, the empty cells were filled by using statistical method, Mode-The most occurred values but it could have been a biased decision because dataset is already imbalanced. Hence at the end, null values from all the columns were dropped. Further it has been observed that dataset values are imbalanced. Video _games has value count of 21727 while musical _instruments has value count 10273. Similarly in the task 2, value counts of 5 ratings is 18706 while for 2 ratings its only 1502. Hence, in order to train model unbiased, Class weight= 'balanced' argument was passed when training models. This argument allows system to treat each values equally and give more weightage to the values which has lesser count. As we are handling the text classification in this machine learning model it is important that the texts are in standardize format. The machine understand capital word 'A' and lowercase 'a' differently. Hence we transformed all the text data into lowercase using string method .lower(). Further the punctuations were removed as well. Sometimes millennial customer's write in Alpha-numeric format like "I l0ve the product". So to minimize such confusions for machines digits were removed from the text columns using RegEx(regular expression) methods. In the next step, Stop words were removed. Stop words are common words like very, some, between etc. To save up the space and speed up the process of prediction, these stop words were removed using NLTK corpus. Further lemmatization process was applied which turns the words into its base formats. Further Label encoder was applied on product _category column to replace categories into numerical format. In order to evaluate the model, Dataset was split into training and testing dataset before feeding it to the ML algorithm model. This process allows us to evaluate the model after training on testing data which was never fed to the algorithm and works like real data. train _test _split method of Sklearn.model _selection was used to split the data.

IV. Methods

For the task1, to predict customer ratings, Multiclass classification algorithm Linear SVC is used with the help of sklearn.linear_library's logistic regression. while for task2, in which we had to identify product category, Binary logistic classification was used. Classification algorithms comes under supervised learning algorithm which provides discrete valued

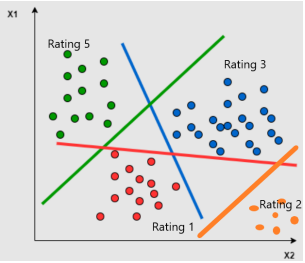

Fig. 1. MultiClass classification using SVM- This figure demonstartes how ratings are divided by the different hyperplanes in SVM algorithm .

output. In our problem statement as we have to classify between product categories and ratings; Binary classification and multi-class classification using SVM was an optimal solution. SVM or Support vector machine divides the output into two groups and find an optimal boundary between the outputs. In case of Multiclass classification SVM algorithm divides the task into multiple binary classification problem.

Before fitting the dataset in Model, Tf-IDF algorithm is used to transform text data into numerical data. TF-IDF stands for Term Frequency Inverse Document frequency. It measures the originality of a word in the whole documents.First it counts the total appearance of a word in particular documents. Further it takes ratio of word count of x to the number of docs the word x appears in. It gives rare term high weightage while common term lower weightage. This allows it to distinguish between the comments precisely in order to identify the sentiment. The arguments passed to the TFID vectorizer are max _df = 0.5, which simply means terms appearing in more than 50% of the documents will be discarded while min _df = 10 means terms appearing in less than 10 documents will be ignored. Further the stop _words = 'English' is given which simply means all the stop words from the english language will be removed. the n _gram values are given (1,5) which will consider 5 sequences of words from a given sample of texts. Then we fit the model in LinearSVC classifier by using Independent variable x(text) and dependent variable y(review score) of training data.After fitting and prediction model was able to perform with 62% accuracy. This accuracy can be increased by manually balancing the data.

Second method used is Binary classification. In binary logistic regression, data is linearly transformed. Then the output is converted into probabilities. Based on the measure of probability, a data point is assigned to particular class. After text preprocessing and fitting the data in TFIDF algorithm, we use GridsearchCV to fine tune the hyper parameters. Grid searchCV loop through predefined hyperparameters and fit the model on training set. Hyperparameter tuning refers to selection of optimal hyperparameters for learning algorithm and minimizing loss function. In the next step, GridsearchCV implemets fit and score method by using logistic regression estimator and c=10 as regualarization value. After implemeting and evaluating the model, Training set accuracy is 98% while Testing test accuracy is 91%.

V. Experiments and evaluation

In the TFIDF vectorizer, the value of n-gram range impacted the model performance. Initially for task1, where we have to predict the rating from 1 to 5, n-gram value was set to (1,3); which gave the accuracy score of 59%. But after changing it to (1,5) the accuracy score raised to 62%. It is concluded that the sequence of words while analysing texts are important for classifier to fit the model and get better accuracy by learning. Because in the binary classification task, value of n-gram were set to (1,3). Because in this task a model had to only predict the category of product. It did not have to analyse the sentiments fully. Even the single keywords like 'game' or 'rhythm' can help model to predict a category. So this is how the sequence of words can impact the output of the model. Further to avoid overfitting of model, the value of c = 20 is passed to Linear SVC model. C refers to the regularization. Regularization reduces the variance in data. Regularization has allowed to increase the accuracy of logistic regression model by minimizing the noise causing data. It is always a good practice to split the data into training and testing set. Now that we have trained our model, we can analyse the accuracy of the models based on the testing set we have kept aside. Firstly, we feed the independent variables of testing data (x _test) to the model. In return model will give an output of predicted values which is saved in y _pred variable. Further, we will compare y _pred and actual y _test values. By using sklearn.metrics library we can compute Accuracy, Precision, Recall and F1 score of the model. Accuracy score is a ratio of sum of True positives and true negatives to the sum of True positives, true negatives, false positives and false negatives. For Task1, the accuracy score is 62% while for task 2 it is 91%. Precision score calculates the ratio of number of True positives(TP) to the number of total positive(Tp+FP). Recall is the ratio of true positive(Tp) to the addition of total positives(Tp) and False Negatives(Fn). F1 score is calculated based on precision and recall. Sklearn.metrics module is used to compute the scores.

TABLE I
METRICS EVALUATION OF RATING(1-5) PREDICTION

ratings	Precison	recall	f1-score
1	0.39	0.36	0.37
2	0.36	0.23	0.28
3	0.40	0.36	0.38
4	0.41	0.40	0.41
5	0.76	81	0.79

TABLE II
METRICS EVALUATION OF PRODUCT CATEGORY PREDICTION

Accuracy	Precison	recall	f1-score
0.918	0.995	0.978	0.986

VI. Discussion and future work

During the implementation of the machine learning model, It is occurred to me that Exploratory data analysis plays a very important role in model performance. It gives a lot of information about the dataset. In order to train the mode, first we have to know about the data we are dealing with. Apart from that, during text classification Data can contain so many symbols, sentiments and punctuations. So to train the machine text classification it is important that the data is clean and follows standardize format.There comes a role of text preprocessing. These are initial steps which make a huge impact in model performance. In the binary classification model of product category prediction task, The model has performed well with good accuracy score. But in the Rating prediction, even after trying various algorithms like KNN, naive-bayes, Random forest classifier the model performance was between 55-60%. SVM algorithm worked well in performance with accuracy score of 62%. In order to increase the model performance in future, we shall deal with the data first. First step will be handling the imbalance. further the suitable algorithm shall be fitted.

VII. Conclusions

In this report, machine learning model based on Binary classification and Multi-class classification on a dataset of Amazon reviews is implemented. It involved major machine learning processes like Exploratory Data Analysis, Data/text preprocessing, Fitting the model into right algorithm, Making predictions and at the end Evaluation of the model using performance metrics.

References

[1] Scikit learn, Model Evaluation: quantifying the quality of predictions, April 2021.

[2] Hucker Clerk Marius, Multiclass-classification with support vector machine, 2020.

[3] Prashant Gupta"Regularization in Machine learning, Nov 2017.